Investigating Expectations for Voice-based and Conversational Argument Search on the Web

Johannes Kiesel
Bauhaus-Universität Weimar
johannes.kiesel@uni-weimar.de

Kevin Lang
Bauhaus-Universität Weimar
kevin.lang@uni-weimar.de

Henning Wachsmuth
Paderborn University
henningw@upb.de

Eva Hornecker
Bauhaus-Universität Weimar
eva.hornecker@uni-weimar.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

ABSTRACT

Millions of arguments are shared on the web. Future information systems will be able to exploit this valuable knowledge source and to retrieve arguments relevant and convincing to our specific need—all with an interface as intuitive as asking your friend "Why . . . ?". Although recent advancements in argument mining, conversational search, and voice recognition have put such systems within reach, many questions remain open, especially on the interface side. In this regard the paper at hand presents the first study of argument search behavior. We conduct an online-survey and a focused user study, putting emphasis on what people *expect argument search to be like*, rather than on what current first-generation systems provide. Our participants expected to use voice-based argument search mostly at home, but also together with others. Moreover, they expect such search systems to provide rich information on retrieved arguments, such as the source, supporting evidence, and background knowledge on entities or events mentioned. In observed interactions with a simulated system we found that the participants adapted their search behavior to different types of tasks, and that up-front categorization of the retrieved arguments is perceived as helpful if this is short. Our findings are directly applicable to the design of argument search systems, not only voice-based ones.

CCS CONCEPTS

• **Human-centered computing** → **Sound-based input / output**; **User studies**; • **Information systems** → **Search interfaces**.

KEYWORDS

conversational search, argument search, voice-based search

ACM Reference Format:
Johannes Kiesel, Kevin Lang, Henning Wachsmuth, Eva Hornecker, and Benno Stein. 2020. Investigating Expectations for Voice-based and Conversational Argument Search on the Web. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343413.3377978

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377978

1 INTRODUCTION

"Why?" is a question we are frequently asked, ask others, and ask ourselves. Why don't you buy an electric car? Why should I vote for her? Why should I not get a duck as a pet? Suitable answers include one or more arguments that are relevant to answer the question, possibly enriched with counter-arguments to a stance intended in the question or other anticipated arguments; e.g.: Ducks eat these annoying snails and do not really need a pond.

Continued advancements in argument mining [6] and voice interfaces [18] may soon allow for similar conversations with machines, but with instant access to a huge number of arguments from a wide variety of views. Unlike formal logic, natural language arguments are usually defeasible [34]. Conceptually, each argument consists of a claim (conveying a stance on a topic) and reasons that support (or attack) the claim. Argument search engines then try to match the user's query with the claims of their indexed arguments and present the respective reasons. While current argument search engines do not support core elements of argumentative discussions—like doubting answers, demanding explanations, or specifying circumstances—progress in conversational search interfaces [21] promises to alleviate such limitations. Conversational argument search may aim to inform the user, to back up the user's stance, or it may follow the *Socratic method*, questioning their stance. Moreover, advances in speech recognition and synthesis have made argument search by voice feasible, allowing for intuitive interaction by leveraging the experience of human-human discussions we engage in daily. However, a discussion of use cases for argument search is missing and argument search systems have mostly been evaluated through topic-coverage and precision [27, 33]. Systems are thus build under an unproven assumption of usefulness.

We here present the first studies on how users expect to, want to, and actually do interact with an argument search system, namely through a voice-based and conversational interface. Specifically, we detail two user studies: an online survey with 500 valid participants on expectations for motivations, situations, and desirable features for voice-based and conversational argument search, and a user study with a simulated system and 18 participants aiming at insights into actual behaviors and perceptions of using the 'system'.

We found that voice-based argument search is expected to (1) be useful to convince others and to come to decisions, but also for entertainment, (2) be used mainly at home, both when alone and with friends, and (3) to provide conversational features, like specific information and reasons on request. While not all argumentative

information needs require a voice-based and conversational interface, such an interface is useful in many situations and can be built on top of display-based ones. Our work focuses on voice-based conversational interfaces and presents results that can be applied directly in their development. But several findings are applicable to display-based argument search as well, conversational or not.

2 BACKGROUND

This paper is the first to combine advancements in the three separate fields of argument mining, conversational search, and voice-based interfaces. The implemented procedure of data collection and analysis is largely inspired by previous voice interaction research.

2.1 Argument Search on the Web

In pioneering work on argument search, Rahwan et al. [22] proposed the World Wide Argument Web: a semantic, web-based, browsable graph of interconnected arguments that was implemented later on [5]. However, separated from the World Wide Web, this web requires volunteers to grow, which limits its size.

The first argument search engine that indexes arguments from web pages was implemented in 2017: Wachsmuth et al. [33] mine and index arguments from debate portals and present pro- and con-arguments for topics entered in a search box.[1] Stab et al. [27], on the other hand, index heterogeneous web pages entirely, and identify arguments at query time using machine learning.[2] The approach by Levy et al. [16] is similar, but no web service is made available. In all these cases, the authors claim to achieve high coverage of relevant topics. These systems thus, to some extent, fulfill the promise of instant access to the wealth of arguments available on the web.

Recently, the first user-centered evaluation of argument search was conducted [20]. The authors evaluated the argument relevance as well as rhetorical, logical, and dialectical quality. However, it remains unclear whether these metrics actually correlate with actual usefulness. Moreover, some use cases for argument search go beyond classical information retrieval evaluation, for example fostering information-literacy [26] or making meaning [24].

Industry also took first steps towards argument search. In 2018, Microsoft announced to answer argumentative queries to their search engine, Bing, with a juxtaposition of web pages in favor and against.[3] We are not aware of any evaluation of this functionality.

The voice-based conversational argument search proposed in this paper would likely be implemented on top of the search engines discussed above. Indeed, most indexing and retrieval pipelines of the conceptualized system would be identical to existing ones, and APIs of search engines could be used to leverage their power.

2.2 Conversational Search

Conversational search is currently widely regarded as one of the most important topics in information retrieval [7], as it allows the user to formulate complex information needs stepwise and intuitively [21]. However, current conversational assistants fail to keep up with this promise, and users thus restrict their usage to simple commands far away from natural language [17]. Indeed, most advances in conversational systems do not directly apply to a search setting, but are integrated in rather specific chat-bot settings [36, 37]. The theory of information seeking dialogues exists for decades [3]. However, research in conversational information seeking still mainly uses mock-ups [32] or observes human dialogues [29, 30] to better understand how the theory could be put into practice. Moreover, current information retrieval metrics require a single result list per query, which does not apply to conversations [13]. On the other hand, general guidelines for interaction design with artificial intelligence systems were developed [1], which can assist in the design of conversational search systems.

Though conversational search is in its infancy, advantages of a conversational interface for argument search are already apparent. Specifically, the five properties of a conversational system [21] are all helpful in argument search: argumentative questions often depend on the context they arise in. Hence, a system that helps the user to express their need (property: user revealment), for example by asking for clarification (mixed initiative) and building knowledge on the user (memory), seems especially useful. Furthermore, since users will only use an argument search engine they trust, it is important that the system communicates its capabilities (system revealment) to avoid failing to live up to the user's expectation [18]. To establish trust, a system should also present counter-arguments, and thus has to reason about argument sets (set retrieval).

2.3 Voice-based Search

Voice-based search is used widely: In 2016, about 20% of all queries from Android phones came via voice.[4] Especially when compared to touch-based search, voice input is much faster [23]. Recently, researchers started to analyze voice-only search for situations without a display [12, 29]. This is particularly relevant for argument search, as arguments are needed throughout daily life. But while voice-enabled smartphones and watches become ubiquitous, keyboards are not. Furthermore, voice interfaces allow for an intuitive interaction as people are used to ask for arguments verbally [18].

However, current voice assistants fail to communicate their capabilities and to live up to user's expectations [17]. Despite decade-long efforts in dialogue research, a generic evaluation framework is missing, and may not be feasible at all [10]. The problem lies in the diversity of situations in which voice-based interfaces may be used. Results obtained in an office (e.g., [28]) might not apply to a domestic or outdoor setting. Evaluation frameworks usually assume a user who is dedicated to the task, but voice-based interfaces allow for quick task-switching and simultaneous other activities. Thus, user satisfaction is insufficient as sole performance yardstick [10].

This work builds on these insights in order to present results on why, when, and how people expect voice-based and conversational argument search to be useful, followed by a detailed study of the actual user behavior in one especially relevant search context.

2.4 User Studies of Voice Assistants

In the field of HCI, a number of studies have emerged on how voice assistants are used in real-life contexts [17, 19, 25]. Audio

[1] https://www.args.me; API: https://www.args.me/api-en.html
[2] http://www.argumentsearch.com/
[3] https://blogs.bing.com/search-quality-insights/february-2018/Toward-a-More-Intelligent-Search-Bing-Multi-Perspective-Answers
[4] https://www.forbes.com/sites/miguelhelft/2016/05/18/inside-sundar-pichais-plan-to-put-ai-everywhere/

If voice-based argument search were available for you today, how likely would you be to use it…
Example setting: You are sitting alone at the breakfast table at home and are reminded by the news about an important election in your country next week. You ask your voice assistant to give pro and con arguments for specific parties to help you make a final voting decision.

	Extremely likely	Very likely	Somewhat likely	Not so likely	Not at all likely	don't know
…alone at home	○	○	○	○	○	○
…to make a voting decision	○	○	○	○	○	○

Figure 1: Examplatory question for the likeliness of using voice-based argument search in specific situations and motivations.

records and logs are used to reveal how current devices are made accountable and embedded into everyday social situations [19] or to identify task categories and temporal patterns [9]. Such analysis, however, raises privacy concerns [11]. Lab studies have found how task/goal types determine factors for user satisfaction (e.g., task completion versus effort) [14]. Wizard of Oz methods are employed to determine how users would naturally interact with speech interfaces, revealing behavior such as the use of implicit conversational cues rather than direct questions [31]. User studies with simulated systems, like in this paper, circumvent current technological limitations but still elicit valuable insight into how users would interact with the next generation of devices, thereby informing their design.

3 SURVEYING USER EXPECTATIONS

So far, a voice-based, conversational argument search system exists as concept only. Many questions are yet to be answered regarding use cases, user profiles, or the presentation of retrieved arguments. We hence conducted an online survey to learn about plausible motivations and situations for using voice-based argument search, as well as expected functionality, of a wide population.

The survey contains questions on participants' background as well as their opinion on motivations, situations, and features for voice-based argument search. In order to familiarize participants with argument search engines, the survey referred them to an example result page of the args.me web service. For the survey, we adopted questions and answers of the widespread product testing and service market research surveys of SurveyMonkey,[5] for which they claim a total usage of more than 100,000 times. The survey design was informed by a local pilot with 6 participants and one that we distributed via mailing lists, reaching 90 participants.

In order to reach a diverse sample, participants were hired via the Mechanical Turk marketplace.[6] We selected the 11 countries with an own Amazon Alexa localization, so that we can assume a general knowledge of voice assistants. Still, we specifically required participants to have heard of voice assistants in the task description. Unfortunately, only five Mexicans took part in our survey, whom we exclude for sake of comparison. Furthermore, we exclude 6 participants who checked the highest or the lowest rating for at least 70% of the questions. Refilling excluded participants, we collected the opinions of 50 unique participants each from Australia, Brazil, Canada, Germany, India, Italy, Japan, UK, and USA. On average, participants required 12 minutes to complete the survey. Following ethical guidelines, we paid participants the average minimum wage of their country. The answers are published in anonymized form.[7]

[5]https://www.surveymonkey.com
[6]https://www.mturk.com/
[7]All resources of this paper are available at https://doi.org/10.5281/zenodo.3490947

Of the 500 final participants, most use voice assistants either frequently (46%) or rarely (45%). Moreover, 1% identified as non-binary, 32% as female, and 67% as male, while most reported to be between 18 to 30 (55%), 31 to 49 (37%), or 50 to 64 (6%) years old.

3.1 Expected Situations and Motivations

In which situations would people utilize voice-based argument search, and with what motivations? Understanding the scope of use is important to inform development of voice-based argument search systems. Voice-based interfaces are useful when typing is inappropriate or impossible, but in some situations their use may feel socially awkward. Moreover, different situations and motivations lead to different kinds of questions. For example, people may ask "why should I …?" rather in private. On the other hand, while discussing with others one might ask a voice-based system to provide (counter-)arguments. Additionally, situation and motivation influence how results should be presented. Is it better to provide an overview of the retrieved arguments, or just a single key argument? Would users attempt to get silly answers by asking silly questions?

In order to better understand in which situations and for which motivations someone would use voice-based argument search, we collected participants' expectations for six exemplary scenarios. Specifically, we asked participants to imagine themselves in some specific situation/location with some specific argumentative information need (cf. Figure 1). For both, they should rate their likeliness of using voice based-argument search on a five-point scale with the option to specify "don't know." Table 1 shows the combinations of motivation, situation, and audience employed in the final survey.

For *situations*, we use the categorization of Efthymiou and Halvey [8] by location and audience. While the original paper distinguishes six location and six audience categories—when combined, resulting in 36 different situations—we merged related location and audience categories (e.g., with partner, family, friends) and dropped edge-cases for argument search (e.g., while driving) to keep the survey length manageable. We decided to use the location categories "at home", "at work", and "in public", as well as audience categories "alone", "with friend", and "with strangers". As some combinations of location and audience would be similar (e.g., work-alone and public-alone) or edge-cases (home-strangers), we used only six of the nine possible combinations to shorten the survey (cf. Table 1).

Figure 2 visualizes how participants of the online survey rated their expected use of voice-based argument search in the selected situations. We highlight three observations:

Observation: *People do not expect to use voice-based argument search in the presence of strangers.* Our participants' expectations corroborate related findings for search on smartwatches [8]: As Figure 2 stresses, many would not use a voice-based system around

Table 1: Combinations of situations and motivations that were employed in the online survey as shown in Figure 1.

Situation	Motivation	Example in survey (Situation/Motivation)
Home, alone	Decide on sth.	Alone at home/make a voting decision
Work, strangers	Convince sb.	In front of customers/convince my colleague
Public, alone	Entertainment	Alone in the park/entertain myself
Work, friend	Convince sb.	At work with a friend/convince my friend
Public, strangers	Decide on sth.	In a crowded store/make a buying decision
Home, friend	Entertainment	At home with a friend/have fun with a friend

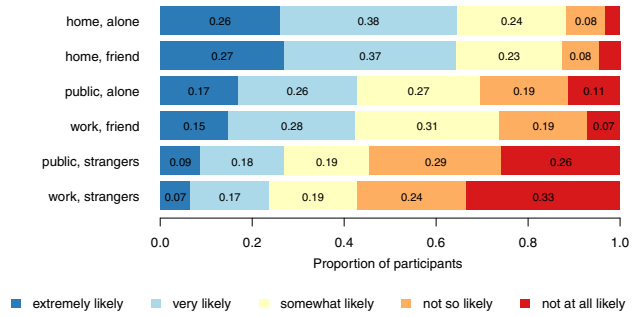

Figure 2: Expected likeliness of using voice-based argument search in different situations.

strangers. In the comments, participants told they would feel embarrassed, afraid of poor answers, or—at work—unprofessional.

Observation: *Most people can imagine using voice-based argument search when at home.* Most participants saw it at least as somewhat likely for them to use a voice-based argument search system at home (Figure 2), either with a friend (87%) or alone (88%). More than 25% even saw their use in such situations as extremely likely, which shows the potential impact of voice-based argument search.

Moreover, for many participants, usage at home is more likely than elsewhere. Controlling for audience, i.e. comparing the ratings for "home-alone" to "public-alone" and for "home-friend" to "work-friend", about 21% of participants prefer both situations at home over the respective other, similar to search on smartwatches [8].

Observation: *Some people prefer using voice-based argument search alone, others with a friend.* A closer inspection of the ratings revealed two groups of participants with distinct preferences: 55% of the participants who rate usage at home alone as extremely likely do rate usage with a friend less likely. On the other hand, a very similar amount (57%) of those who rate usage at home with a friend as extremely likely, rate usage alone less likely. Therefore, some participants would use voice-based argument search when alone, about the same amount of participants would rather use it together with others, and yet another group would use it in both situations. At first glance, this seems to contradict results from a study on search on smartwatches, where usage alone was clearly prefered over when together with friends [8]. But different to that study, our scenarios contain situations where the participants would use the voice assistant together with their friend, not just in their presence.

We categorized *motivations* according to a respective categorization of argumentative texts as persuasive (i.e., changing the stance of someone else) or deliberative (i.e., weighing options of a decision)

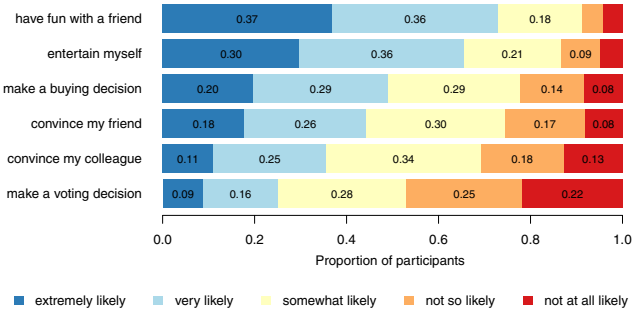

Figure 3: Expected likeliness of using voice-based argument search for different motivations.

[34]. We thus distinguish between convincing somebody and deciding on something. Moreover, as fun is often part of voice-based interfaces [10, 17, 18] and has a place in online discussion forums,[8] we include entertainment as a third category of motivation. Figure 3 shows how participants of the online survey rated their expected likeliness of use of voice-based argument search for the selected motivations. We highlight two observations:

Observation: *People expect voice-based argument search to be entertaining.* Most participants rated their use of voice-based argument search for the two scenarios with entertainment motivations as extremely or very likely (73% and 66%; cf. Figure 3). Moreover, 91% of participants see the use case to have fun with a friend at least somewhat likely. Entertainment thus seems a major factor attracting users to voice-based argument search systems.

In addition, 179 participants (36%) specified their use for at least one entertainment case as strictly likelier than all non-entertainment cases. Therefore, it seems that entertainment scenarios can attract people to an argument search system that would otherwise not use it, perhaps because they would mistrust the technology.

Observation: *People can imagine using voice-based argument search both for making a decision and for convincing others.* Using a voice-based argument search system in order to make a voting decision was seen as likely by the least participants. Indeed 22 participants gave a comment that they would mistrust the software and question the bias of the algorithm or company. For comparison, only one participant said they would fear to be spied upon in their voting decision. Nevertheless, although many regard voting as an identity-defining decision (in the comments, participants call it "very personal", "mine at all cost", or "mine alone"), some can imagine to use it as one of several sources of information (specifically mentioned in 9 comments, especially "to get initial info" or "an initial overview"). Therefore, the participants rated a use for this decision still more likely than a use at work with a stranger (cf. Figure 2). Thus many participants expect such a system to be useful for basic argumentation tasks. However, they clearly see a difference in what to decide on or whom to convince: Figure 3 shows that more participants would use such a system for buying decisions (which most participants probably use the Internet for already) than for voting decisions, and more for convincing a friend (who is presumably more forgiving) than for convincing a colleague.

[8] E.g., http://www.debate.org/opinions/funny/

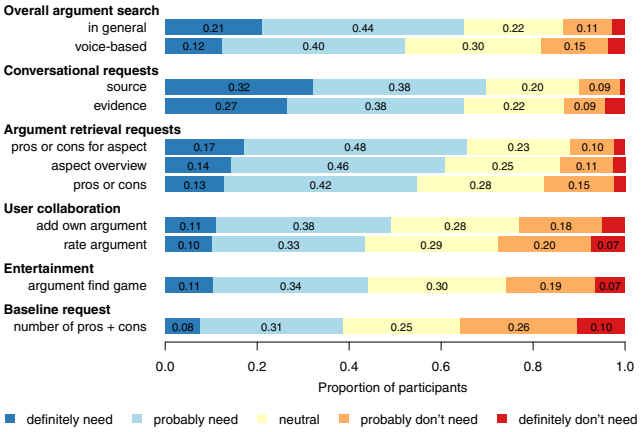

Figure 4: Appreciation of possible features of a voice-based argument search engine. Feature categories added in bold.

3.2 Expected Functionality

Which functionality would people expect or desire a voice-based argument search system to have? Answers to this are useful to determine key functionality (red route tasks). While additional features are usually not negative—especially in a voice-based interface where they would not "clutter" the interface—development time is limited and developers have to focus on core features.

To better understand which functionality people expect or desire in a voice-based argument search system, we collected participants' expectations of how much they would need nine features and argument search in general in their life. The features include a barely helpful baseline request, three core argument retrieval requests, two conversational requests that inquire on arguments listed beforehand, two actions to collaborate on improving the system, and an argument search game for entertainment (cf. Figure 4). For the baseline request, we chose giving the absolute number of pro and con arguments for a topic. Even though participants were instructed to assume flawless implementation, a total number does not provide much insight given arguments have different relevance. All features should be rated on five-point scale, again with the option of "don't know". Additionally, the survey asked participants to specify other needed features. Argument search in general is seen as something they definitely or probably need by 65% of participants. The fraction is a bit less for voice-based argument search (52%), but still shows a rather large interest. We highlight four additional observations:

Observation: *People expect to converse with voice-based argument search systems on arguments.* More than half of participants state to definitely or probably need the standard retrieval of pro and con arguments that current argument search engines focus on (e.g., [27, 33]), but about two third do so for a deeper investigation of retrieved arguments: getting evidence for, or source of the argument, which one participant even calls "essential for credibility." For voice-based argument search systems, where each information element transmitted to the user requires them to listen for longer, such information can only reasonably be provided on a per-request basis. This requires a conversational system that keeps track of told arguments to which the user could refer to in their next request.

Observation: *People expect to be able to specify their information need precisely.* Another feature that is—albeit slightly—seen as needed by more participants than the standard retrieval of pro and con arguments is to do the same retrieval for specific aspects of a topic. In contrast to the standard retrieval that can be used to gain an overview of the most important arguments on a topic, specifying an aspect would allow receiving more focused results. An aspect overview is thus seen as needed by a similar amount of participants. The focused results are especially useful if one already has an overview of the arguments related to a topic and now wants to dig deeper or explore controversial aspects. A focused search is also helpful if the user wants to find an argument again. Indeed, the idea of the aspect overview stems from one of the participants in the pre-study that was distributed via mailing lists.

In this regard, several participants suggested additional ways to state their information need more precisely. Suggestions included restricting results to arguments from neutral or scientific sources, to certain views or even a specific person. The participants thus see a clear need for argument search to allow for precise questions. Voice-based interfaces are especially suited for this, as they allow for such specifications in natural language and thus without the need to learn a specific query syntax or interface switches (cf. Section 2).

Observation: *People would appreciate being able to contribute, but see such functionality not as essential.* Being able as a user to contribute to the search index was a bit less appreciated than the standard retrieval of pro and con arguments. Motivations for contribution are not only altruistic: user ratings help the search engine to provide argument lists that are more tailored to the user by employing recommendation technology. This would especially be relevant for buying decisions (cf. Section 3.1). However, while there is some appreciation, it is clearly not the focus of attention for users.

Observation: *People would prefer the usual argument search for entertainment over an argument find game.* While entertainment can be a motivation for about 90% of participants to use argument search for (cf. Figure 3), only 35% of the participants would say they need the option to play an argument game.

3.3 Expected Argument Ranking Factors

How should the retrieved arguments be ranked? Ranking is a core mechanism of search engines and arguments allow for a variety of ranking factors that need to be weighed against each other. In fact, ranking gets even more important for voice-based interfaces, as skimming results takes much more time than with visual interfaces. While survey results cannot be expected to be directly translatable into a ranking function, it is nevertheless useful to understand users' expectations for the design of such a function.

In order to understand which factors are expected to influence the argument result ranking, we asked participants to rate six factors on a 6-point Likert scale from most to least important. The 6-point scale allowed to specify a total ordering, but this was not enforced. The factors were selected to both cover a variety of possible influences and contrast different ways to achieve a similar goal. They included argument strength (determined by machine learning or by user ratings), coverage (by aspects or by argument sources, e.g., forum posts and editorials), source reliability, and recency. Figure 5 illustrates the ratings. We highlight two observations:

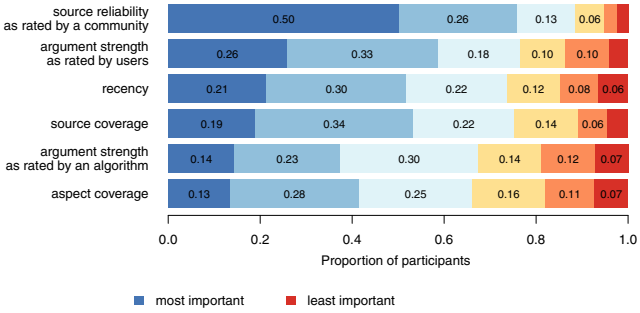

Figure 5: Desired importance of factors for ranking arguments in the result list.

Observation: *People expect reliable arguments.* The visually most striking observation from Figure 5 is that half of all participants value source reliability as most important. This result is impressive, especially when considering that the suggested reliability measure could be flawed: our suggestions derive reliability from the rating of some (not further described) community. For comparison, a factor with similar foundation, namely argument strength as rated by users, was rated as much less important. We thus conclude that argument reliability is the top concern for users.

Observation: *People trust other users over algorithms.* While 26% of participants saw argument strength as rated by users as most important, only 14% did so for the same measure as computed by an algorithm. This is surprising, given we asked participants to assume a flawless algorithm for strength prediction. Explanations might thus be that many participants were confused by the term of argument strength or that there is a general mistrust to strength ratings provided by algorithms.

4 OBSERVING USER INTERACTIONS

While online surveys are well-suited to quickly collect opinions and expectations, user studies that involve a working system provide much more realistic and detailed data [18]. However, due to their high time cost, user studies are usually limited to a few specific use cases. In our user study presented below we focus on the situation "alone at home", which is, according to the online survey, one of the preferred situations (see Section 3.1). We wanted to learn how users would interact with a voice-only and conversational argument search system, i.e., what kind of information is requested, how the information is requested, and how users react to system actions.

In addition to observing the interactions with the system, we wanted to analyze whether argument retrieval is done differently if the task is to convince others versus for own decision making, or, when it is beneficial or not to present an aspect overview. We hypothesized that participants who seek arguments to convince others focus more on arguments supporting their own opinion than do participants who seek arguments for private decision making. We thus created pairs of search tasks whose topic is identical within the pair, but whose motivation differs within the pair (see Table 4). We also picked up a suggestion from the online survey, namely, to first provide an *overview* of the aspects of the retrieved arguments along with the option to choose a subset that is then presented.

Since no such system exists at the moment, we simulated it with a human in the loop who interacted voice-only with the participant, who was seated in a separate room. This setup resembles Wizard-of-Oz studies [32], except the fact that our participants were aware of the simulation. To ensure consistency the simulation was always done by the same person, and a detailed behavior guideline for the human simulator was created up-front and extended during a pilot study with two participants and the main study as necessary. All interactions were recorded and transcribed: The instructor presented tasks to the participants, but then left them alone and sat next to the simulator during interactions in order to take notes and to ensure consistency. Participants were informed about the data collection and gave consent for its usage for research. This procedure was approved by our University's privacy officer.

The study started with a privacy declaration, basic demographic questions similar to those of the online survey, and an instruction sheet. To bring all participants to a similar level, the instruction sheet—besides giving an overview of the study structure—listed example information they could ask for (e.g., arguments, argument categories, further evidence, sources) as well as general voice assistant commands (e.g., "stop" and "help"). Additionally, it stated that answers by the system may be made-up (since the information asked might not be readily available). To avoid that participants cling to these examples, we removed the sheet before the tasks.

Participants had to complete four argument finding tasks illustrated by a corresponding scenario each, namely one for each combination of decide/convince and with/without aspect overview. As some participants in the pilot had problems to imagine certain scenarios, we offered six to pick from: in a first round, they selected two from three tasks (buy an electric car, visit the zoo, study abroad), and in the second round, two from three further tasks (stop eating meat, support rights for conscious AI, support school uniforms).

Tasks were selected from the online discussion platform Kialo.[9] We carefully selected tasks that are plausible for both decide and convince motivations and that provide a reasonable number of arguments and aspects. For half the participants, tasks in the first round were phrased so as to decide for themselves and tasks in the second round so as to convince a friend, whereas for the other half of participants it was the opposite order (counterbalancing). The aspect overview was provided only in the second task of a round so that participants already experienced the "usual" system when rating the overview. Participants were instructed that they had 10 minutes to collect arguments until they were satisfied, but no participant took longer than 8 minutes. The shortest of such conversations took 96 seconds and is transcribed in Table 2 (a).

After each task, participants answered a short questionnaire about their experience and—if the aspect overview was presented in the task—their choice of aspects. Table 2 (b) shows an example of the aspect overview in a conversation. To assess the aspect overview, they rated the system helpfulness, speed, pleasantness of use, predictability, naturalness of conversation, and the structuredness of responses for the preceding task on a 5-point Likert scale. To understand their use of the aspect overview, the questionnaire also asked for their reasoning behind selection of aspects. Finally, a post-study questionnaire asked for general feedback on the system.

[9] https://www.kialo.com

18 participants completed the study. Since the study required presence, these were recruited via mailing lists of our university, which led to a younger sample: 72% were 18–30, and 28% were 31–49 years old. Less participants reported a frequent (17%) or rare (39%) use of voice assistants than in the survey, so nearly half of participants (44%) had not used one beforehand. The gender distribution has been similar though, with 67% male and 33% female participants. Like for the survey, the questionnaires and data—including the coded transcript—are available publicly at the same place.

4.1 Interactions with the Simulated System

What information would users request from a voice-based argument search system? How will they converse with the system? Where the expectations collected in Section 3.2 target the value that users see in features, the interactions observed in the user study provide an idea of how often they would end up using such features.

To gain a quantitative overview of how participants interacted with the simulated system, we categorized each turn of both participant and system. We utilized the categories employed by Azzopardi et al. [2], but merged categories that were very similar for argument search or occurred rarely. The categories used are: *Generic* (not related to search), *Navigating* (specify (participant) or acknowledge (system) a topic or aspect), *Inquiring* (query for (participant) or offer (system) more arguments or information on a specific argument), and *Revealing* (provide feedback (user) or present arguments or information (system)). While the inquire and reveal actions can be approximately mapped to the QRFA-model of information-seeking dialogues [30], generic and navigate actions fall outside of the scope of that model as they are not directly related to gathering information. Table 2 shows example excerpts of some of the conversations with utterances are marked with the respective category.

Table 3 provides an overview of interactions in terms of number of actions per category and speaker. Note that there are far more system actions than user actions as the system usually answered a request and directly followed up with a suggestion of how to continue the conversation. We here focus on the actions of participants, as these are the relevant ones for our analysis.

As Table 3 shows, about two third of the participant's actions were of information seeking nature. About a third of all actions of the participants were inquiries for certain information. On inspection, we find that most of these are requests for arguments (219 actions, 64% of all inquiring actions of the participants). Unexpectedly, they also often inquired for context information such as definitions, product information, comparisons, statistics and research data (65 actions, 19%). Another 29 actions inquired a specific aspect. However, only 20 actions (6%) inquired for evidence and only 4 actions (1%) inquired the source. For illustration, Table 2 (c) shows an excerpt of a conversation in which the participant inquired on both the source and further evidence for an argument that was just retrieved for them. This is in stark contrast to the high appreciation of these features in the survey (cf. Section 3.2). Another third were revealing actions, where the participant either confirmed or negated wanting to hear from a specific category, from pro or con, or more retrieved arguments. The remaining third were either generic actions to keep the conversation alive (14%) or navigating actions to change the state of the system (22%).

Table 2: Transcripts of (a) the shortest conversation, (b) a category overview, and (c) an inquiry for source and evidence. Labels denote speakers (Participant or System) and action categories (Generic, Navigating, Inquiring, or Revealing).

P [Alexa.]$_G$
S [Hello.]$_G$
P [Hi,]$_G$ [is introducing school uniform helps students achieve better grads?]$_N$
S [Open Topic *Students should wear School Uniform.*]$_N$ [There are in total 11 pro arguments and 11 con arguments for this topic.]$_R$ [Which one do you want to hear?]$_I$
P [Pros, please.]$_I$
S [First, having all students wear the same uniform helps create a sense that you belong somewhere and maintain a good school either by culture or spirit. Second, it makes all the children at a school equal no matter what their family background or income is. Third, introducing uniform can reduce crime in schools, especially violence and theft, partly due to better discipline, but also because students no longer come to school wearing desirable designer clothes.]$_R$ [Do you want to hear more arguments?]$_I$
P [Can you please tell me some of the cons?]$_I$
S [First, people who are forced to wear certain items, such as a head scarf, to school will be pulled out of school. Second, rich kids can still show expensive designer labels with watches or smartphones. Third, Uniforms can actually lead to more violence, as they make students from rival schools much more obvious in the street, which can increase bullying and fighting.]$_R$ [Do you want to hear more arguments?]$_I$
P [Oh, no thanks.]$_R$ [That was helpful.]$_N$
S [Good Bye.]$_G$
P [Bye.]$_G$

(a)

P [Alexa, can you give me some arguments for school uniforms?]$_N$
S [Open Topic *Students should wear School Uniforms.*]$_N$ [There are in total 5 categories related to this topic.]$_R$ [Do you want to hear arguments related to *Feeling of Belonging*?]$_I$
P [Are there more topics?]$_I$
S [There are in total 5 categories.]$_R$
P [Can you please name me the categories?]$_I$
S [The categories are: *Feeling of Belonging*, *Social Inequalities*, *School Achievement*, *Crime and Violence* and *Expense.*]$_R$ [Do you want to hear arguments related to *Feeling of Belonging*?]$_I$
P [Uhm. Yes.]$_R$
S [Do you want to hear arguments related to *Social Inequalities*?]$_I$
...
S [Do you want to hear arguments related to *Expense*?]$_I$
P [No.]$_R$
S [There are in total 2 pro and 3 con arguments for the category *Feeling of Belonging.*]$_R$ ...

(b)

S ... [First, electric cars fight global warming better than hybrids.]$_R$ ... [Do you want to hear con arguments?]$_I$
P [No,]$_R$ [but I would like to hear the source of the first argument.]$_I$
S [The argument is originated from the article of the portal *faz.net.*]$_R$
P [Do you have any scientific research regarding this topic.]$_I$
S [Do you want to hear evidence for this argument?]$_I$
P [Yes.]$_R$
S [Evidence that speaks for this argument is the study of Dr. Prof. Walter Herbernikus in 2014, which showed that electric cars have a two times better impact on global warming than hybrids.]$_R$
P [Okay. Thank you.]$_N$

(c)

Table 3: Categories of actions and amount of times the participants and the (simulated) system performed actions of that category in the user study. Percentages are relative to all actions by the participant and all actions by the system. Table 2 provides examples for each category transcribed from conversations of the study.

Action categories		Number of actions	
Name	Examples from category	Participant	System
Generic	Ask to/provide confirm(ation), help, repeat	143 (14%)	149 (8%)
Navigating	Specify/use topic or aspect	232 (22%)	200 (11%)
Inquiring	Query for/offer pros, cons, more, source	312 (30%)	622 (34%)
Revealing	Provide answers, feedback, information	356 (34%)	842 (46%)
Total		1043 (100%)	1813 (100%)

Table 4: Total number of participant-requests for pro and con arguments for the 12 tasks of the user study.

Task		Number of requests		
Motivation	Topic	Pro	Con	$\sum$
Decide whether to	buy an electric car	28 (60%)	19 (40%)	47
Convince a friend to		30 (65%)	16 (35%)	46
Decide whether to	not visit the zoo	18 (50%)	18 (50%)	36
Convince a friend to		18 (72%)	7 (28%)	25
Decide whether to	study abroad	5 (38%)	8 (62%)	13
Convince a friend to		7 (41%)	10 (59%)	17
Decide whether to	stop eating meat	9 (41%)	13 (59%)	22
Convince a friend to		9 (53%)	8 (47%)	17
Decide whether to	support rights for conscious AI	7 (44%)	9 (56%)	16
Convince a friend to		12 (52%)	11 (48%)	23
Decide whether to	support school uniforms	15 (54%)	13 (46%)	28
Convince a friend to		17 (57%)	13 (43%)	30

4.2 Effects of Task Motivation

Is there a generic argument search behaviour or do people adapt to a specific motivation? If they adapt, the search system could use this to identify the motivation and in turn also adapt its own behavior. For example, if the user looks for arguments to convince someone, the system might want to return especially convincing arguments to the user and focus less on coverage of sources.

Table 4 shows the number of pro and contra argument requests for each of the twelve tasks of the study, grouped by topic. While percentages vary significantly between topics, for all six topics, the percentage of pro argument requests is higher if the task description stated a motivation to convince a friend (of the pro-side) than when it stated that participants want to make a decision. This observation indicates that participants indeed adapt their search behavior to motivation. However, the large variance between topics might make it difficult to detect such changed behavior: where the pro-side was requested between 38% and 60% when deciding for one self, the increase when convincing others can be as small as 3%.

4.3 Effects of the Aspect Overview

How to present users the arguments they need quickly? While vision-based systems can display several arguments in parallel and thus allow for fast skimming, voice-based systems can only speak one argument at a time. Methods to identify and skip irrelevant arguments are thus of great interest for voice-based systems. For our user study, we picked up the suggestion of a survey participant and provide an overview of all aspects before the actual argument result list (cf. Table 2b). The system would ask the participant for each aspect of the topic whether the user would be interested, and—in case not—remove the respective arguments from the result.

For a quantitative assessment, we compare ratings from the post-task questionnaire between tasks with and those without the aspect overview. The ratings were given on a five-point Likert scale and map to $[1; 5]$ (best to worst). This allows to compute the average rating, μ, as well as the standard deviation, σ. Since our data is ordinal and of pairwise observations, we resort to a one-sided Wilcoxon signed-rank test [35] for significance testing.

While participants rated the system with aspect-overview as significantly faster ($p = 0.006$; $\mu = 2.11, \sigma = 1.01$ vs. $\mu = 2.33, \sigma = 1.01$) and more pleasant to use ($p = 0.001$; $\mu = 2.17, \sigma = 1.08$ vs. $\mu = 2.31, \sigma = 1.09$) than without, they did not rate it as more helpful ($p = 0.156$; $\mu = 2.14, \sigma = 0.93$ vs. $\mu = 2.19, \sigma = 1.01$), more behaving as expected ($p = 0.055$; $\mu = 2.11, \sigma = 1.01$ vs. $\mu = 2.22, \sigma = 1.07$), more natural ($p = 0.097$; $\mu = 2.36, \sigma = 1.20$ vs. $\mu = 2.53, \sigma = 1.16$), or more structured ($p = 0.821$; $\mu = 1.97, \sigma = 0.97$ vs. $\mu = 1.89, \sigma = 0.75$). Furthermore, they would not recommend the system with aspect overview more than the system without ($p = 0.092$; $\mu = 2.47, \sigma = 1.08$ vs. $\mu = 2.44, \sigma = 1.03$). Overall, differences in ratings are rather small.

However, on closer inspection of participants' textual feedback, we found that the aspect overview was useful for some, but annoying for others. Five participants were confused because the arguments were not directly presented after selecting an aspect, but only after all aspects were either selected or dismissed. Two participants rated the overview as not helpful when there are only few aspects, and thus less need for an overview (one topic had only two aspects) and found it overwhelming if there were many aspects (as much as nine). This feedback reveals major problems in our design of the aspect overview, but also shows ways to handle these.

5 DISCUSSION

The observations from the online survey and user study provide several noteworthy implications for the design of argument search engines. Many implications apply—to some extent—not only to voice-based, but also to display-based and multimodal argument search (e.g., voice input mixed with visual and audio output on smartphones). However, the following limitations apply.

5.1 Limitations

First, it is important to consider the time of this work: we expect that voice-based interfaces, conversational search, as well as argument search will develop considerably within the next few years. Thus the questions in the online survey had to be rather generic so as to not depend on currently available technology. However, in case the social acceptability of voice search changes (which is likely, similar to how it has become quite normal to do phone calls in public), the perceived convenience in different situations will change too. Moreover, there is currently no system for voice-based and conversational argument search, and thus survey participants had to rely on their imagination. To ease this, we pointed them to a functional argument search system and employed scenarios from

everyday life with a clear motivation of use. Pilot studies indicated that participants could easily imagine themselves in these scenarios. For the user study, we tried to cope with this by simulating the system. Moreover, humans have argued for ages, with the principles observed by Aristotle still being applicable. Especially for voice-based interfaces—which support and focus on natural interaction (cf. Section 2)—predictions could be quite accurate.

A limitation specific to the user study is that it considers only the situation home, alone. While this is the situation most participants of the survey considered as convenient (cf. Section 3.1), other situations are plausible. Clearly, interactions will be different when the device is used in the company or even in collaboration with others (cf. [19]). However, we had to leave this up for future research.

5.2 Implications

A voice-based interface is considered as useful for all argument search tasks that Section 3.1 discusses: more than half of the survey participants stated for every motivation it to be somewhat likely that they would use such an interface for it. However, we observed a tendency towards considering argument search for less "critical" use cases, such as fun discussions and convincing friends. A natural explanation is the still unproven quality and maturity of argument search in general. As soon as search results turn out to be useful and reliable, we expect people to become more open to other motivations. In this regard, our analysis revealed that fun discussions may help to motivate people to try out argument search. However, providing support for fun discussions is double-edged, since building trust in retrieval results is decisive for the success of an argument search engine but difficult when being funny.

Currently, people expect voice-based argument search to be useful mainly in private situations (Section 3.1). However, unlike in a related study [8], our participants could easily imagine using such a system when being in company of a friend. It may then become necessary to feature speaker identification in order to allow for a joint interaction with the system. However, this setting provides many new challenges and opportunities for a conversational search system, because the system would need to model several users along with both their individual and their shared knowledge.

Section 3.2 showed that the source of arguments is seen as extremely important, and that an argument search engine should provide and consider such information in its ranking. However, users relatively rarely ask for the source when observed in an actual task scenario (Section 4.1). The expected importance of the source relates with the current public discussion on what is usually referred to as "fake news": misinformation spread by dubious web pages that present themselves as news publishers. The low number of observed requests for an argument source may seem at odds with the highly-rated importance. However, users of an argument search engine would not check the source for arguments as long as they intuitively judge these as plausible, which might be the vast majority. Furthermore, our experimental setup might have caused users to trust the "system" more than usual, given that they knew it was prepared for the topic. Still, the ability to check the source of counter-intuitive arguments may remain very important, even if only used sporadically. Moreover, one could argue that the system itself should notify the user if they may not trust the source.

Most importantly, as the system learns which sources the user distrusts, should it exclude those from its results? Recent research speaks in favor of such an approach: It has been found that people trust sources more that share their own political attitude, even if the source argues against the party line in that case [4, 15]. If, on the other hand, a system has not adapted to the user's distrust—as for new users or ones that do not want the system to store data on them, thereby preventing adaptation—it remains to be investigated how an objective reliability judgment may be achieved.

Besides the importance of source reliability, several other implications can be derived from the ratings of ranking factors. The strength of arguments—but only if judged by humans—is seen as another very important criterium, which indicates a distrust in algorithmic ranking. Moreover, recency of arguments is seen as "most important" by 21% of the survey's participants. However, to the best of our knowledge, the creation or publication time of an argument is not even discussed as a possible factor in the argument search literature. Furthermore, while source and aspect coverage are not the primary factors, they are still seen as most important by 19% and 13%, which suggests the use of result list diversification algorithms that already exist for traditional web search.

A finding with design implications for all conversational argument search systems is that the participants adjusted their requests to the motivation (deciding for themselves vs. convincing a friend). The effect is probably not limited to difference in the relative number of pro and con requests shown in this paper, though. Nevertheless, a detection of user motivation and adaptation of the conversational systems seems thus feasible and reasonable.

Lastly, as our analysis of both desired functionality and actual behavior showed, users expect several features from a voice-based argument search engine that suggest a conversational interface. Desired functionality includes checking the retrieved arguments and filtering results, whereas the observed behavior showed that methods to provide deeper information like encyclopedic definitions, product information, or statistics would be well-appreciated. While these features seem particularly useful for voice-based search, they will likely benefit users of a display-based interface, too.

6 CONCLUSION

In recent years, research has produced several advancements that put human discussions with machines into reach. While there are still problems to solve—some identified in this work—the presented paper analyzes how such discussions *are expected to look like* and hence, (1) what designers of such systems need to pay attention to and (2) which directions require further research.

Based on data from a forward-looking online survey with 500 participants from 10 countries and a focused user study with 18 participants, we have formulated 11 observations, centered around the questions why, when, and how people expect to perform (voice-based) argument search. We found that voice-based argument search is seen as useful for all argumentative information needs, in particular, when at home alone or together with friends. Furthermore, it should be provided by a conversational interface for interactive search refinement, it should retrieve and incorporate context information for the arguments at hand and, not at least, it should especially focus on arguments from reliable sources.

REFERENCES

[1] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proc. of CHI*. 3:1–3:13. https://doi.org/10.1145/3290605.3300233

[2] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *Proc. of CAIR*. 8.

[3] Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications* 9, 3 (1995), 379–395. https://doi.org/10.1016/0957-4174(95)00011-W

[4] Adam J. Berinsky. 2017. Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science* 47, 2 (2017), 241–262.

[5] Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. 2013. Implementing the argument web. *Commun. of the ACM* 56, 10 (2013), 66–73.

[6] Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *Proc. of IJCAI*. 5427–5433.

[7] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL). *SIGIR Forum* 52, 1 (2018), 34–90.

[8] Christos Efthymiou and Martin Halvey. 2016. Evaluating the Social Acceptability of Voice Based Smartwatch Search. In *Proc. of AIRS*. 267–278.

[9] David Graus, Paul N. Bennett, Ryen W. White, and Eric Horvitz. 2016. Analyzing and Predicting Task Reminders. In *Proc. of UMAP*. 7–15.

[10] Kristiina Jokinen and Michael F. McTear. 2009. *Spoken Dialogue Systems*. Morgan & Claypool Publishers.

[11] Joseph Jofish Kaye, Joel Fischer, Jason I. Hong, Frank R. Bentley, Cosmin Munteanu, Alexis Hiniker, Janice Y. Tsai, and Tawfiq Ammari. 2018. Panel: Voice Assistants, UX Design and Research. In *Extended Abstracts of CHI*. 5.

[12] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *Proc. of SIGIR*. 1257–1260. https://dl.acm.org/citation.cfm?id=3209978.3210160

[13] Julia Kiseleva and Maarten de Rijke. 2017. Evaluating Personal Assistants on Mobile devices. *CoRR* abs/1706.04524 (2017), 4.

[14] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proc. of CHIIR*. 121–130.

[15] David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. Shorenstein Center, https://shorensteincenter.org/combating-fake-news-agenda-for-research/. Accessed: 2018-09-24.

[16] Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an Argumentative Content Search Engine using Weak Supervision. In *Proc. of COLING*. 2066–2081.

[17] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proc. of CHI*. 5286–5297.

[18] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences* (1st ed.). O'Reilly Media, Inc.

[19] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proc. of CHI*. 640:1–640:12.

[20] Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument Search: Assessing Argument Relevance. In *Proc. of SIGIR*. 4.

[21] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proc. of CHIIR*. 117–126.

[22] Iyad Rahwan, Fouad Zablith, and Chris Reed. 2007. Laying the foundations for a World Wide Argument Web. *Artif. Intell.* 171, 10-15 (2007), 897–921.

[23] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Y. Ng, and James A. Landay. 2017. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *IMWUT* 1, 4 (2017), 159:1–159:23.

[24] Ian Ruthven. 2019. Making Meaning: A Focus for Information Interactions Research. In *Proc. of CHIIR*. 163–171. https://doi.org/10.1145/3295750.3298938

[25] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proc. of DIS*. 857–868.

[26] Catherine L. Smith and Soo Young Rieh. 2019. Knowledge-Context in Search Systems: Toward Information-Literate Actions. In *Proc. of CHIIR*. 55–62. https://doi.org/10.1145/3295750.3298940

[27] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proc. of NAACL-HTL*. 21–25.

[28] Johanne R. Trippas, Damiano Spina, Falk Scholer, Ahmed Hassan Awadallah, Peter Bailey, Paul N. Bennett, Ryen W. White, Jonathan Liono, Yongli Ren, Flora D. Salim, and Mark Sanderson. 2019. Learning About Work Tasks to Inform Intelligent Assistant Design. In *Proc. of CHIIR*. 5–14. https://doi.org/10.1145/3295750.3298934

[29] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2017. How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis. In *Proc. of CHIIR*. 325–328.

[30] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A Data-Driven Model of Information-Seeking Dialogues. In *Proc. of ECIR*. 541–557. https://doi.org/10.1007/978-3-030-15712-8_35

[31] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proc. of CHI*. 208:1–208:7.

[32] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proc. of CHI*. 2187–2193.

[33] Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proc. of ArgMining*. 49–59.

[34] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

[35] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83.

[36] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proc. of ACL*. 2204–2213.

[37] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. In *Proc. of ACL*. 1128–1137.